

Szacunek liczby cudzoziemców w Polsce

Maciej Beręsewicz

Katedra Statystyki, Uniwersytet Ekonomiczny w Poznaniu
Ośrodek Metodologii Badań Ludnościowych, Urząd Statystyczny w Poznaniu

 BERENZ / ncn-foreigners / ojalab
 /  @mberesewicz
mberesewicz.bsky.social

Mieszkaniec, czyli kto? Warszawa 24-25 marca 2025



POZNAŃ UNIVERSITY
OF ECONOMICS
AND BUSINESS

Plan prezentacji

- 1 Motywacja
- 2 Probabilistyczna integracja danych
- 3 Długość pobytu populacji cudzoziemców w Polsce
- 4 Wybrane wyniki
 - Jakość rejestrów administracyjnych
 - Wybrane wyniki
- 5 Podsumowanie

Plan prezentacji

- 1 Motywacja
- 2 Probabilistyczna integracja danych
- 3 Długość pobytu populacji cudzoziemców w Polsce
- 4 Wybrane wyniki
- 5 Podsumowanie

[O nas](#)[Aktualności](#)[Stanowiska](#)[Z naszych miast](#)[Publikacje](#)[Wydarzenia](#)[Metropolie w liczbach](#)[Projekty](#)

[Home](#) / [Czy wiemy, ile nas jest? Nowy raport Fundacji im. Stefana Batorego i Unii Metropolii Polskich](#)

Rysunek 1: Raport Fundacji im. Stefana Batorego i Unii Metropolii Polskich (źródło: https://metropolie.pl/fileadmin/news/2023/06/Czy-wiemy-ile-nas-jest_raport.pdf)

Polska Prezydencja w Grupie Roboczej ds. Statystyki (CWPS)

[Strona
główna](#)[Aktualności](#)[Program](#)[Kalendarz](#)[Spotkanie Wysokiego
Szczepła](#)

**Polska Prezydencja w Grupie Roboczej
ds. Statystyki (CWPS)**

Wybrane aspekty jakości rejestrów administracyjnych

Tabela 1: Liczba rekordów dot. populacji cudzoziemców według rejestrów zgodnie ze stanem na 31.12.2023

Rejestr	Liczba rekordów	Ukraina		Brak adresu [%]
		Ogółem	Status UKR	
NFZ	5 371 434	3 168 914	953 979	34,7
PESEL	4 260 492	3 137 978	954 005	92,1
KEP 2022	1 796 867	1 120 414	53 935	9,3
ZUS	1 206 357	832 781	236 715	5,95
PESEL (RDK)	1 155 032	995 222	429 459	89,3
UdSC	701 830	363 085	5 543	9,0
POLON	242 491	111 334	11 323	0,00
KEP (2023; DG)	86 200	49 596	13 791	1,4
KRUS	34 587	31 918	2 045	0,01

Wybrane aspekty jakości rejestrów administracyjnych

Tabela 2: Liczba rekordów według rejestrów zgodnie ze stanem na 31.12.2023 dołączonych do rejestru PESEL na podstawie identyfikatora PESEL

Liczba rejestrów	Liczba cudzoziemców	%
1 rejestr (PESEL)	10 182	0,2
2 rejestrów	1 677 504	39,4
3 rejestrów	1 520 732	35,7
4 rejestrów	592 899	13,9
5 rejestrów	332 708	7,8
6 rejestrów	110 980	2,6
7 rejestrów	151 197	0,4
8 rejestrów	290	0,0

Cel prezentacji

- Omówienie problematyki jakości rejestrów administracyjnych w kontekście populacji cudzoziemców: integracja bez kluczy, problemy braków danych, niespójności czy ustalania miejsca pobytu.
- Przedstawienie probabilistycznego blokowania rekordów na potrzeby integracji danych bez kluczy.
- Przedstawienie wyników w zakresie określania populacji cudzoziemców w zakresie długości pobytu na podstawie zintegrowanych rejestrów administracyjnych (ZUS za okres listopad 2022–grudzień 2023).

Plan prezentacji

- 1 Motywacja
- 2 Probabilistyczna integracja danych
- 3 Długość pobytu populacji cudzoziemców w Polsce
- 4 Wybrane wyniki
- 5 Podsumowanie

Przykład z polskich rejestrów administracyjnych

Tabela 3: Przykładowe dane z rejestrów

Imię	Drugie imię	Nazwisko	Data urodzenia
Miguel	Luis	Pereira Tinoco	1969
Miguel Luis		Pereira-Tinoco	1969
Miguel		Pereira-Tinoco	1968
Thi	Hy	Dao	1993
Dao Thi Hy			1993
Thi		Dao	1992

Idea łączenia danych bez kluczy

A THEORY FOR RECORD LINKAGE

IVAN P. FELLEGI AND ALAN B. SUNTER

Dominion Bureau of Statistics

A mathematical model is developed to provide a theoretical framework for a computer-oriented solution to the problem of recognizing those records in two files which represent identical persons, objects or events (said to be *matched*).

A comparison is to be made between the recorded characteristics and values in two records (one from each file) and a decision made as to whether or not the members of the comparison-pair represent the same person or event, or whether there is insufficient evidence to justify either of these decisions at stipulated levels of error. These three decisions are referred to as *link* (A_1), a *non-link* (A_2), and a *possible link* (A_3). The

Idea łączenia danych bez kluczy

SCIENCE ADVANCES | REVIEW

COMPUTER SCIENCE

(Almost) all of entity resolution

Olivier Binette¹ and Rebecca C. Steorts^{2,3*}

Whether the goal is to estimate the number of people that live in a congressional district, to estimate the number of individuals that have died in an armed conflict, or to disambiguate individual authors using bibliographic data, all these applications have a common theme—integrating information from multiple sources. Before such questions can be answered, databases must be cleaned and integrated in a systematic and accurate way, commonly known as structured entity resolution (record linkage or deduplication). Here, we review motivational applications and seminal papers that have led to the growth of this area. We review modern probabilistic and Bayesian methods in statistics, computer science, machine learning, database management, economics, political science, and other disciplines that are used throughout industry and academia in applications such as human rights, official statistics, medicine, and citation networks, among others. Last, we discuss current research topics of practical importance.

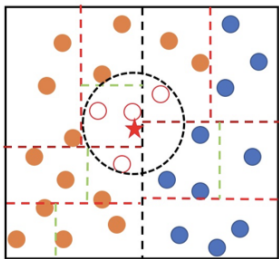
Copyright © 2022
The Authors, some
rights reserved;
exclusive licensee
American Association
for the Advancement
of Science. No claim to
original U.S. Government
Works. Distributed
under a Creative
Commons Attribution
License 4.0 (CC BY).

Algorytmy wyszukiwania przybliżonych sąsiadów

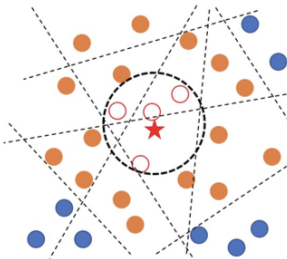
Nearest neighbor search (NNS), as a form of *proximity search*, is the optimization problem of finding the point in a given set that is closest (or most similar) to a given point. Closeness is typically expressed in terms of a dissimilarity function: the less similar the objects, the larger the function values (Wikipedia 2024).

- Cel: kompromis między dokładnością a szybkością.

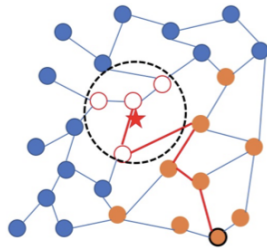
Idea algorytmów ANN algorithms



(a)



(b)

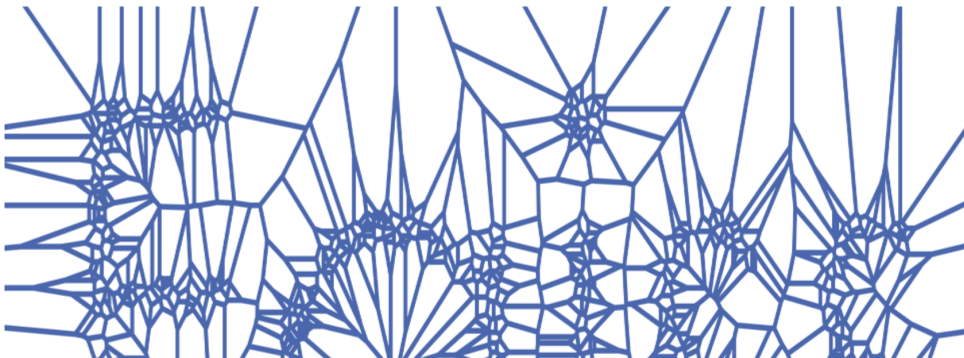


(c)

Rysunek 2: Types of algorithms: a) box, 2) hyperplanes, 3) graphs. Source: Fu, C., Xiang, C., Wang, C., & Cai, D. (2017). Fast approximate nearest neighbor search with the navigating spreading-out graph

O jakich algorytmach mówimy?

Faiss: A library for efficient similarity search



O jakich algorytmach mówimy?



VOYAGER

License [Apache 2.0](#)

Documentation [on github.io](#)

platforms [macOS](#) | [Windows](#) | [Linux](#)

Apple Silicon [macOS and Linux](#)

 Compile, Test, and Deploy [passing](#)

Voyager is a library for performing fast approximate nearest-neighbor searches on an in-memory collection of vectors.

Voyager features bindings to both Python and Java, with feature parity and index compatibility between both languages. It uses the HNSW algorithm, based on [the open-source hnswlib package](#), with numerous features added for convenience and speed. Voyager is used extensively in production at Spotify, and is queried hundreds of millions of times per day to power numerous user-facing features.

O jakich algorytmach mówimy?

rnndescent

An R package for finding approximate nearest neighbors, translated from the Python package [PyNNDescent](#) written by the great Leland McInnes. As the name suggests, it uses the Nearest Neighbor Descent method ([Dong et al., 2011](#)), but also makes use of Random Partition Trees ([Dasgupta and Freund, 2008](#)) as well as ideas from [FANNG](#) and [NGT](#).

You can use rnndescent for:

- optimizing an initial set of nearest neighbors, e.g. those generated by [RcppAnnoy](#) or [Rcpp-HNSW](#).
- using this package for nearest neighbor search all on its own...
- ... including finding nearest neighbors on sparse data, which most other packages in the R ecosystem cannot do.
- and a much larger number of metrics than most other packages.

Zaproponowana procedura

- ➊ Przetwórz dane tekstowe, aby usunąć spacje, przecinki itp.
- ➋ Utwórz q-gramy i przechowuj dane w macierzy rzadkiej
 - {ma, ac, ci, ie, ej}
 - {mac, aci, cie, iej}
- ➌ Utwórz indeks przy użyciu określonego algorytmu
- ➍ Przeszukaj indeks używając tych samych (deduplikacja) lub nowych danych (łączenie rekordów)
- ➎ Utwórz graf do grupowania jednostek
- ➏ **Klastry – bloki do deduplikacji/łączenia rekordów**

Oprogramowanie

blocking 0.1.0 Reference Articles ▾ Changelog

Overview

Warning!

The package is still being developed, so the API and features may change.

Description

This R package is designed to block records for data deduplication and record linkage (also known as entity resolution) using [approximate nearest neighbours algorithms \(ANN\)](#) and graphs (via the `igraph` package).

It supports the following R packages that bind to specific ANN algorithms:

- [rnnodescent](#) (default, very powerful, supports sparse matrices),
- [RcppINSW](#) (powerful but does not support sparse matrices),
- [RcppAnnoy](#),
- [mlpack](#) (see `mlpack::lsh` and `mlpack::knn`).

🏠 / Welcome to BlockingPy's Documentation

[View page source](#)

Welcome to BlockingPy's Documentation

license MIT repo status Active python 3.10+ coverage unknown
pypi v0.1.11 Ruff Run Tests passing last commit february

Contents

- [Getting Started](#)
 - [Installation](#)
 - [Quick Start](#)

Oprogramowanie dostępna na Github: <https://github.com/ncn-foreigners> (autorzy: Tymoteusz Strojny, Maciej Beręsewicz).

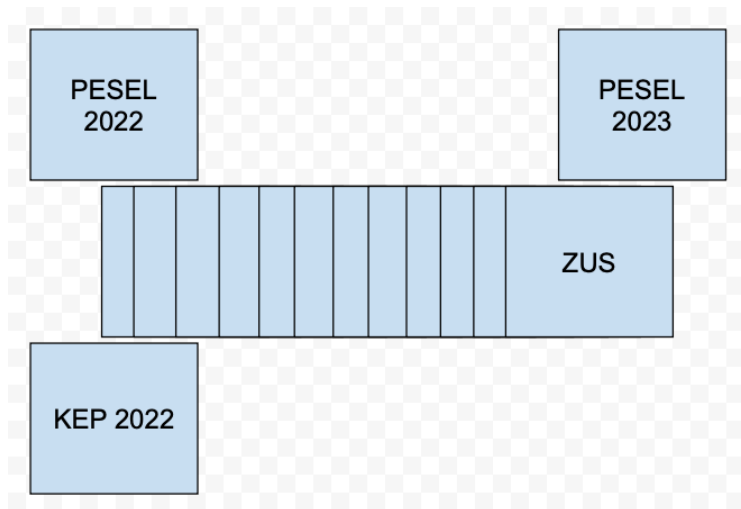
Plan prezentacji

- 1 Motywacja
- 2 Probabilistyczna integracja danych
- 3 Długość pobytu populacji cudzoziemców w Polsce
- 4 Wybrane wyniki
- 5 Podsumowanie

Populacja cudzoziemców wg rejestru ZUS

- W celu ustalenia populacji cudzoziemców oraz długości pobytu w Polsce, w tym Wielkopolsce i Poznaniu przeprowadzono pilotażowe badania wykorzystujące rejestr ZUS, PESEL oraz KEP (MF).
- Zintegrowano dane z ZUS za okres **listopad 2022 – grudzień 2023** oraz rejestr PESEL (stan na 31.12.2022 i 31.12.2023) i KEP (stan na 31.12.2022; pełnych danych za 2023 jeszcze nie mamy).
- Skupiono się na ustaleniu dolnej granicy populacji cudzoziemców na podstawie rejestru ZUS (osoby zgłoszone do ubezpieczenia emerytalnego i rentowe oraz prowadzące własną działalność).

Poglądowy schemat



Plan prezentacji

- 1 Motywacja
- 2 Probabilistyczna integracja danych
- 3 Długość pobytu populacji cudzoziemców w Polsce
- 4 Wybrane wyniki
 - Jakość rejestrów administracyjnych
 - Wybrane wyniki
- 5 Podsumowanie

Długość pobytu wg rejestru PESEL?

Tabela 4: Integracja danych z PESEL (osoby żyjące) wg stanu 2022.12.31 i 2023.12.31

PESEL 2022	PESEL 2023	
	NIE	TAK
NIE	–	472 914
TAK	17 836	3 787 578
Razem	17 836	4 260 492

Wynik sugerowałby, że blisko 90% cudzoziemców obecnych w rejestrze PESEL (osób żyjących) przebywało w Polsce 12 miesięcy i więcej. Przy czym większość nie ma wskazanego adresu (oprócz zameldowanych).

Tabela 5: Status adresu w rejestrze PESEL

Status adresu	2022	2023
Brak adresu	3 560 960	4 004 258
Występuje	244 454	256 234

Długość pobytu wg rejestru KEP?

- KEP – Krajowej Ewidencji Podatników
- W momencie przygotowywania prezentacji nie miałem dostępu do danych z 2023.
- Informacje o danych z 2022:
 - Liczba osób (?): 1 796 867, z czego:
 - 22% (ok. 400 tys.) nie ma podanego identyfikatora PESEL.
 - 16% (ok 300 tys.) nie ma podanego adresu.

Długość pobytu wg rejestru ZUS?

Tabela 6: Podstawowe informacje z rejestru ZUS w zakresie populacji cudzoziemców (w tys.

Miesiąc	Rekordy	Osoby	Ma PESEL	Ma adres
2022-11	1 174	1 072	886	560
2022-12	1 183	1 068	886	556
2023-01	1 178	1 049	897	548
2023-02	1 184	1 057	921	551
2023-03	–	–	–	–
2023-04	–	–	–	–
2023-05	1 205	1 075	941	558
2023-06	1 207	1 078	946	557
2023-07	1 223	1 090	961	562
2023-08	1 219	1 089	963	561
2023-09	1 235	1 099	975	563
2023-10	1 218	1 110	985	569
2023-11	1 219	1 111	992	573
2023-12	1 206	1 105	992	569

Przykład historii gminy zamieszkania wg rejestru ZUS

Data	ID=1	ID=7	ID=10
2022-11	–	3021084	1421021
2022-12	–	1421021	–
2023-01	–	1421021	–
2023-02	2468011	1421021	–
2023-03	?	?	?
2023-04	?	?	?
2023-05	2477011	3064049	–
2023-06	2477011	3064049	–
2023-07	2472011	–	–
2023-08	2417042	–	–
2023-09	2417042	2462011	1421021
2023-10	2462011	2201024	1421021
2023-11	2402044	2201024	–
2023-12	–	2204011	–

Dane z ZUS po deduplikacji, probabilistycznej integracji, imputacji

Tabela 7: Dane z rejestru ZUS po procesie "czyszczenia" (w tys.)

Miesiąc	Liczba osób	Braki w adresie	% braków
2022-11	1 027	45	4.17
2022-12	1 025	42	3.97
2023-01	1 017	38	3.64
2023-02	1 019	38	3.55
2023-03	?	?	?
2023-04	?	?	?
2023-05	1 037	38	3.51
2023-06	1 039	38	3.56
2023-07	1 051	39	3.58
2023-08	1 050	39	3.59
2023-09	1 060	39	3.57
2023-10	1 066	40	3.58
2023-11	1 070	40	3.62
2023-12	1 065	40	3.62

Długość pobytu – liczba miesięcy obserwowania w rejestrze ZUS

Tabela 8: Liczba cudzoziemców 18-70 wg liczby miesięcy występowania w zintegrowanych rejestrach (wg stanu na koniec 2023)

Liczba miesięcy	Liczba cudzoziemców	% ogółu
1	386 861	20.359862
2	165 488	8.709363
3	103 266	5.434721
4	102 391	5.388671
5	73 912	3.889868
6	65 249	3.433948
7	63 992	3.367794
8	86 167	4.534828
9	61 534	3.238434
10	65 545	3.449526
11	76 746	4.039017
12	421 347	22.174804
13	227 618	11.979163

Długość pobytu wg stanu na koniec 2023

Tabela 9: Długość pobytu populacji 18-70 na podstawie zintegrowanych danych (stan na koniec 2023)

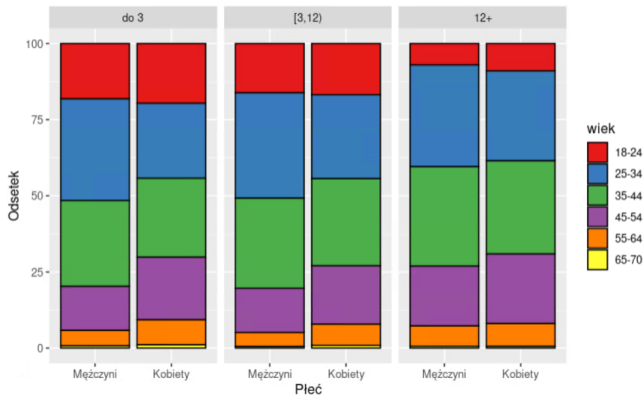
Długość pobytu	Liczba	%
do 3 miesięcy	102 940	9.4
[3, 12) miesięcy	222 295	20.3
12+ miesięcy	769 352	70.3
Ogółem	1 094 587	100

Długość pobytu: Ukraina vs pozostałe

Tabela 10: Długość pobytu populacji 18-70 na podstawie zintegrowanych danych według kraju obywatelstwa (stan na koniec 2023)

Ukraina	Liczba osób	% w grupie
<i>do 3 miesięcy</i>		
Tak	53 758	52.2
Nie	49 182	47.8
<i>[3, 12) miesięcy</i>		
Tak	134 843	60.7
Nie	87 452	39.3
<i>12 i więcej miesięcy</i>		
Tak	559 997	72.8
Nie	209 355	27.2

Długość pobytu: Ukraina vs pozostałe



Rysunek 3: Struktura populacji cudzoziemców 18-70 według długości pobytu, płci i wieku (wg stanu na koniec 2023)

Plan prezentacji

- 1 Motywacja
- 2 Probabilistyczna integracja danych
- 3 Długość pobytu populacji cudzoziemców w Polsce
- 4 Wybrane wyniki
- 5 Podsumowanie**

Podsumowanie

- Przedstawione wyniki należy traktować jako **dolną granicę** wielkości populacji (nie uwzględniono dzieci i osób poza rejestrem ZUS).
- Źmudna procedura przetwarzania i czyszczenia, integracji danych bez kluczy (opracowaliśmy własne procedury w języku R).
- Wyzwania: opracowanie metodyki śladów życia, długości pobytu i miejsca zamieszkania, populacja dzienna i nocna.
- Dyskusja: zmiana sposobu liczenia populacji – *fractional counting* – uwzględniamy niepewność i nie liczymy "głów"(ang. *head count*), a prawdopodobieństwo przebywania (np. 80% Poznań, 20 % Radojewo).
- Prace badawcze wspierane przez NCN OPUS 20 (2020/39/B/HS4/00941).
- Narzędzia open source: BlockingPy, blocking i nonprobsvy:
<https://github.com/ncn-foreigners>.

Dziękuję za uwagę!